A Novel Approach for Predicting User Behavior for Improving Web Performance

Priyanka Makkar[1], Payal Gulati[2], Dr. A.K. Sharma[3]
Department of Computer Science & Engg[1], Department of Computer Engg[2,3].
Manav Rachna College of Engineering[1], YMCA University of Science & Technology[2,3]
Faridabad, Haryana[1,2,3]
preyankamakker@gmail.com, gulatipayal@yahoo.co.in, ashkokkale2@rediffmail.com

Abstract*— The rapid growth in the amount of information and the number of users has lead to difficulty in providing effective search services for the web users and increased web latency; resulting in decreased web performance. Although web performance can be improved by caching, the benefit of using it is rather limited owing to filling the cache with documents without any prior knowledge. Web prefetching becomes an attractive solution wherein forthcoming page accesses of a client are predicted, based on access log information. This paper proposes a novel approach for increasing web performance by analyzing and predicting user behavior both by collaborating information from user access log and website structure repository.

Keywords-*:* **web log file, pattern discovery, petri nets, web site structure repository, prefetching**

I. INTRODUCTION

The size of publically indexed World Wide Web has probably surpassed 24.39 billion pages in June 2010 [2] and as yet growth shows no sign of leveling off. more information becomes available on web it is more difficult to provide effective search services for Internet users thus increasing web latency. User perceived latency comes from several sources such as bandwidth, speed, overhead etc. However network latency cannot be removed due to the latency caused by propagation delay, but can be minimized. In accordance of "*Eight Second Rule*" [1], it can be deduced that user care about web latency and lot of efforts are taken to minimize the latency perceived by the user. Caching of web documents at various points in the network (client, proxy, server) [3,4] has been developed to reduce the latency. Although web caching has already widely used in the www, the benefit from web caching are becoming limited due to the rapid changes of network resources, while web prefetching can optimize the www in many respects [5]. This motivates this work, in this paper prefetching is done so as to improve web performance. Prefetching is the (cache initiated) speculative retrieval of a resource into a cache based on user access log; in the anticipation that it can be served from cache in the future [6] leading to reduction in web latency. This work proposes a novel approach for increasing web performance by analyzing user behavior both by collaborating information from user acesss log and website structure repository.

This paper is organized in the following way. Section II discusses related work done in this domain. Architecture of the proposed system is presented in Section III. Sequence of steps of the work is discussed in Section IV. Finally Section V comprise of the conclusion.

II. RELATED WORK

Padmanabhan et al [6] investigated ways of reducing retrieval latency. A scheme called prefetching was introduced in which clients in collaboration with servers prefetch web page that the user is likely to access soon, while he/she is viewing the currently displayed page. L Fan et al [7] investigate an approach to reduce web latency by prefetching between caching, proxies, and browsers. Research on predictive Web prefetching has involved the important issue of log file processing and the determination of user transactions (sessions) from it [8, 21]. Chen, Cooley and Pie [9, 10 and 11] provide various data mining algorithms for the path traversal patterns and how to efficiently mine the access patterns from the web logs. Pirolli and Pitkov [12] predict the next web page by discovering the longest repeating subsequence in the web sessions. Liu et al [13] used association rules for web access predicions. Yang et al [14] studied different association rule based methods for web request prediction. Using association rules for web access prediction involves dealing with too many rules and it is not easy to find a suitable subset of rules to make accurate and reliable predictions. Padbanabham and Mogul [15] use N-hop Markov models predicted the next web page users will most likely access by matching the user's current access sequence with the user's historical web access sequences for improving prefetching strategies for web caches. Sarukkai [16] used first-order Markov models to model the sequence of pages requested by a user for predicting the next page accessed. Liu et al [17] integrated association rules and clustering for reducing the overhead associated with large database. Cadez et al [18] integrated clustering and first order

Markov model to increase accuracy. Kim et al [19] combined association rules and Markov model for web access prediction. In this paper, data mining techniques are applied for extracting user access patterns from web access logs. Futher this work uses application of petri nets (PN) for enhancing web usage mining. Web structure is extracted using parsing algorithm, from which incidence matrix is built. The web structure information in the incidence matrix and the reachability properties obtained from the PN model help in path competition process.

III. PROPOSED WORK

This section describes the architecture of the proposed system shown in Figure 1. Following subsection describes various components of the proposed system.

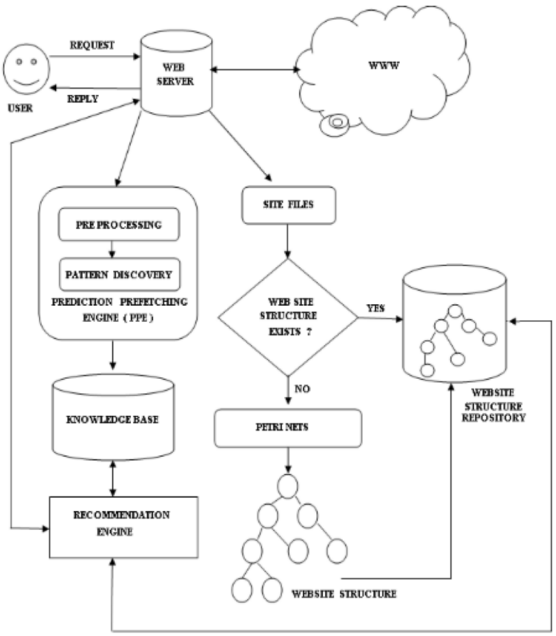

Figure 1. Proposed System Architecture

A. *Web Server*

The primary function of the web server is to deliver web pages to the users. The records of all the users/clients that send requests to the server are kept on the web logs, which reside at the server side. This log file helps in analyzing user access patterns and in predicting next page likely to be accessed by the web user.

B. *Prediction Prefetching Engine (PPE)*

Prediction Prefetching Engine (PPE) processes the past references, tracking the user behavior to deduce the probability of future access of the web page based on access log information. It comprises of two parts preprocessing and pattern discovery phase, discussed in following subsections.

- **Pre preprocessing** Data preprocessing involves data cleaning, user and session identification. Cleaning the web log file involves removal of irrelevant items like the image file (GIF and JPEG) and java script files (JS) etc., as these do not contribute for the patterns relevance. Next step to cleaning is user and session identification. A user session is defined as the sequence of requests made by the single end user during a visit to a particular site. Within a single session, a user may follow links to several pages that belong to the similar pattern but during the same session, it may also be possible that the user might visit some other pages that do not belong to the same pattern i.e. user session may contain the documents belonging to patterns while others that do not and the documents are interleaved in the session. After all the preprocessing is done, the cleaner version of the log is formed called Data mart. Data mart acts as a database on which various data mining operations operate for generating the rules.

- **Pattern Discovery** Data mining is the process of extracting hidden patterns from data. After the preprocessing of the web server log file, data mining techniques are applied. In this work, combination of clustering and association rule mining is proposed. The clustering of web user sessions is done so as to cluster user with similar behavior together. Further association rules mining is applied on clustered sets. Association Rule mining are a major pattern discovery technique. The statistics computed for the association rules, support and confidence are given in equation (i) and (ii).

$$\text{Support } (X \Rightarrow Y) = P(X \cup Y) \quad \text{... (i)}$$
$$\text{Confidence } (X \Rightarrow Y) = P(Y \mid X) \quad \text{.... (ii)}$$

C. *Knowledge Base*

The knowledge base is a repository of extracted rules which have been derived by applying data mining techniques. A knowledge base containing rules is shown in Figure 2.

R1: A → B
R2 : A → B →C
……
……
……
……
R3 : B →D

Figure 2. Sample Knowledge base

D. Petri nets

Petri Nets (PN) is a high-level graphical model widely used in modeling system activities with concurrency.PN can store the analyzed results in a matrix for future follow-up analyses. According to the definition it is formally defined as a 5-tuple PN of (P, T, I, O, M0), where

(1) P = {p1, p2… pm}, a finite set of places;
(2) T = {t1, t2… tn}, a finite set of transitions; P∪ T≠Ø, and P∩T= Ø;
(3) I: P x T→N, an input function that defines directed arcs from places to transitions, where N is a set of nonnegative integers;
(4) O: T x P →N, an output function that defines directed arcs from transitions to places;
(5) M0: P →N, the initial marking. A marking is an assignment of tokens to a place;

PN is carried out by firing transitions. A transition, t, is said to be enabled if each input place, p, of t contains at least an amount of token equal to the weight of the directed arc connected to t from p. In a PN model, we can utilize the different token amounts in the places to represent the different system states. Since a fire of transition in the system often can be associated with a change of the token amount in a place, PN hence can represent, or model, the system dynamic behaviors via the fire of transitions. An incidence matrix records all token-amount changes in all places after all fired transitions. For PN with n transitions and m places, the incidence matrix A, where A=[a_{ij}], is an n×m matrix of integers; its typical entry is given by

$a_{ij} = a_{ij}^{+} - a_{ij}^{-}$
where a_{ij}^{+}= O (t_i, p_j), the weight of the arc from Transition i to its Output Place j, and
a_{ij}-=I (t_i, p_j), the weight of the arc to Transition i from its Input Place j;
a_{ij}^{+}, a_{ij}^{-} and aij represent the number of tokens removed, added, and changed in
Place p_j, respectively, when Transition t_i fires once.

During the processing or operations, this method can also simultaneously trace out what are the possible intermediate states during the transitions from the initial state to the destination one. In a PN model, a marking M_i is said to be reachable from a marking, M_0, if there exist a sequence of transition firings which can transform a marking, M_0, to Mi [20].

In the proposed work data mining techniques are applied for extracting user access patterns from web access logs. Futher this work uses application of petri nets (PN) for enhancing web usage mining. Web structure is extracted using parsing algorithm, from which incidence matrix is built. The web structure information in the incidence matrix and the reachability properties obtained from the PN model help in path completion process.

E. Website Structure Repository

The website structure is composed of the different web pages of the website and navigation within those pages. Web structure repository consists of the site structures of all the websites residing at the server.

F. Recommendation Engine

Whenever user request for URL, PPE sends the URL to the recommendation engine, which in turn does prediction based on rules from knowledge base and hints from site structure repository. It also does path completion process. Finally based on this, it prefetch the web page in clients cache before he/she explicitly request for, thus decreasing web latency and improving web performance.

IV. ALGORITHM

This Section discuses the sequence of steps of the work done by the proposed system.

Algorithm: UpdateWhenNewQuery(q_{i+1})
begin
$qlog_{i+1}$ ← $qlog_i$ U q_{i+1}
for each new query q_{i+1}

RecomendationEngine(q_{i+1}); //based on KB and websiteStrRepository will make recommendations

Preprocess ($qlog_{i+1}$)
DiscoverPatterns($qlog_{i+1}$) // will return rules
KB ← rules //knowledge base
end for
end

Algorithm : RecomendationEngine(q_{i+1})
begin
for each new query
make predictions based on KB

if (Site Structure exists)
break;
else
petrinets(URL, Sitestructure)
websiteStrRepository ← <URL,Sitestructure>
end if

PathCompletion based on site structure
Make predictions based on websiteStrRepository
Prefetch Pages
end for
end

V. CONCLUSION

This paper proposes a novel approach for predicting user behavior for improving web performance. In this prediction and prefetching is done both by collaborating information from user access log and website structure repository. This work overcomes the limitation of path completion. Application of Petri Nets for extracting web site structure helps in path completion process, better prediction, decreasing web latency and improving web performance.

REFERENCES

[1] http://en.wikipedia.org/wiki/Network_performance
[2] http://www.worldwidewebsize.com/index.php?lang=NL
[3] C. Aggarwal, J. Wolf, P. S. Yu. "Caching on the World Wide Web" , IEEE Transactions on knowledge and Data Engineering , Vol. 11, No.1, pp. 95-107, Feb. 1999.
[4] P.Cao and S. Irani. "Cost – Aware WWW Proxy Caching Algorithms", Proceedings 1997 USENIX Symposium on Internet Technology and Systems (USITS'97), pp 193- 206, Jan.1997.
[5] Lei Shi, Yingjie Han, Xiaoguang Ding, Lin Wei, Zhimin Gu. " SPN Model for Web Prefetching and Caching ", Proceedings of the First International Conference on Semantic, knowledge, and Grid (SKG 2005), IEEE, 2006.
[6] Venkata N. Padmanbhan. "Improving World Wide Web Latency", Technical Report, Computer Science Division, University of California, Berkeley, CA, May, 1995.
[7] Fan L., Cao P., and Jacobson Q., "Web prefetching between Low-Bandwidth Clients and proxies: potential and performance." In Proceedings of the Joint International Conference on Measurement and Modeling of Computer Systems., May 1999.
[8] P. Atzeni, G. Mecca, and P. Merialdo, "To Weave the Web," Proc. 23rd Conf. Very Large Data Bases (VLDB '97), pp. 206-215, Aug.1997.
[9] M.S. Chen, J.S. Park, and P.S. Yu, "Efficient Data Mining for Path Traversal Patterns," *IEEE Trans. Knowledge and Eng., Vol.10,* n0.2, pp.209-221, Apr. 1998.
[10] R. Cooley, B. Mobasher and J.Srivastva, "Data Preparation for mining World Wide Web Browsing Patterns,"*Knowledge and information Systems (KAIS), Vol. 1, no.1,* pp. 5-32, Feb. 1999.
[11] J. Pei, J.Han, B. Mortazavi, and H. Zhu, " Mining Access Patterns efficiently from Web Logs," *Proc. Pacific - Asia Conf. Knowledge discovery and Data Mining (PAKDD'00),* Apr. 2000.
[12] Pitkov, J. and Piroli, P. "Mining Longest repeating Subsequences to predict world wide web surfing, Proc. USENIX Symp. On Internet Technologies and Systems, 1999.
[13] Liu, B., Hsu, W., and Ma, Y., "Integrating Classification and Association Mining", Proc. of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD-98), 1998.
[14] Yang, Q., Li, T., Wang, K., "Building Association Rules Based Sequential Classifiers for Web Document Prediction", journal of Data Mining and Knowledge Discovery, Netherland: Kluwer Academic Publisher, 2004.
[15] V. Padmanabhan and J. Mogul, "Using Predictive prefetching to improve World Wide Web latency", *ACM SIGCOMM Computer Comm. Rev., Vol. 26,no.3,* July 1996.
[16] Sarukkai, R.R., "Link prediction and path analysis using Markov chain", proc. of the 9th International World Wide Web Conference on Computer networks, 2000.
[17] Liu, F. Lu Z. " Mining Association rules using clustering", Intelligent Data Analysis, 2001.
[18] Cadez I., Heckerman D., MeekC. Symth P., and Whire S., " Visualization of Navigation Patterns on a website using Model Based Clustering", March, 2002.
[19] Kim, D., Adam, N. Alturi, V., Bieber, M. & Yesha, Y. "A clickstream - based collaborative filtering personalization model: Towards a better performance", WIDM, 2004.
[20] Po-Zung Chen, Chu-Hao Sun, Shih-Yang Yang, "Modeling and Analysis the Web Structure Using Stochastic Timed Petri Nets", Journal of Software, Vol. 3, No. 8, November 2008.
[21] Payal Gulati, A.K. Sharma, Amit Goel, Jyoti Pandey, "A Novel Approach for Determining Next Page Access," icetet, pp.1109-1113, 2008 First International Conference on Emerging Trends in Engineering and Technology, 2008.

AUTHORS PROFILE

Priyanka Makkar received B.E. degree in Computer Science & Engineering with Hons. from Maharshi Dayanand University in 2007 and is pursuing M.Tech. in Information Technology. Presently, she is working as a Lecturer in Computer Engineering department in Manav Rachna College of Engineering, Faridabad. Her areas of interests are Web Mining and Search Engines.

Payal Gulati received B.E.(IT), M.Tech (Computer Engineering) degrees with Hons. from Maharshi Dayanand University in 2006 and 2008 respectively. Presently, she is working as an Assistant Professor in Computer Engineering Department in YMCA University of Science & Technology. She is also pursuing Ph.D in Computer Engineering and her areas of interest are Web Mining, Search Engines and Crawlers.

Prof. A. K. Sharma received his M.Tech. (Computer Science & Technology) with Hons. from University of Roorkee in the year 1989 and Ph.D (Fuzzy Expert Systems) from JMI, New Delhi in the year 2000. From July 1992 to April 2002, he served as Assistant Professor and became Professor in Computer Engg. at YMCA University of Science & Technology, Faridabad in April 2002. He obtained his second Ph.D. in IT from IIIT & M, Gwalior in the year 2004. His research interests include Fuzzy Systems, Object Oriented Programming, Knowledge Representation and Internet Technologies.